

Theory-guided Data Science: A New Paradigm for Scientific Discovery

Anuj Karpatne, Gowtham Atluri, James Faghmous, Michael Steinbach, Arindam Banerjee,
Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar

Abstract—Data science models, although successful in a number of commercial domains, have had limited applicability in scientific problems involving complex physical phenomena. Theory-guided data science (TGDS) is an emerging paradigm that aims to leverage the wealth of scientific knowledge for improving the effectiveness of data science models in enabling scientific discovery. The overarching vision of TGDS is to introduce scientific consistency as an essential component for learning generalizable models. Further, by producing scientifically interpretable models, TGDS aims to advance our scientific understanding by discovering novel domain insights. Indeed, the paradigm of TGDS has started to gain prominence in a number of scientific disciplines such as turbulence modeling, material discovery, quantum chemistry, bio-medical science, bio-marker discovery, climate science, and hydrology. In this paper, we formally conceptualize the paradigm of TGDS and present a taxonomy of research themes in TGDS. We describe several approaches for integrating domain knowledge in different research themes using illustrative examples from different disciplines. We also highlight some of the promising avenues of novel research for realizing the full potential of theory-guided data science.

Index Terms—Data science, knowledge discovery, domain knowledge, scientific theory, physical consistency, interpretability

1 INTRODUCTION

From satellites in space to wearable computing devices and from credit card transactions to electronic health-care records, the deluge of data [1], [2], [3] has pervaded every walk of life. Our ability to collect, store, and access large volumes of information is accelerating at unprecedented rates with better sensor technologies, more powerful computing platforms, and greater on-line connectivity. With the growing size of data, there has been a simultaneous revolution in the computational and statistical methods for processing and analyzing data, collectively referred to as the field of data science. These advances have made long-lasting impacts on the way we sense, communicate, and make decisions [4], a trend that is only expected to grow in the foreseeable future. Indeed, the start of twenty-first century may well be remembered in history as the “golden age of data science.”

Apart from transforming commercial industries such as retail and advertising, data science is also beginning to play an important role in advancing scientific discovery. Historically, science has progressed by first generating hypotheses (or theories) and then collecting data to confirm or refute these hypotheses. However, in the big data era, ample data, which is being continuously collected without a specific theory or hypothesis in mind, offers further opportunity for discovering new knowledge. Indeed, data science is starting to gain prominence in several scientific disciplines, where its role is shifting from providing simple analysis tools (e.g. de-

tecting particles in Large Hadron Collider experiments [5], [6]) to providing full-fledged knowledge discovery frameworks (e.g. in bio-informatics [7] and climate science [8], [9]). Based on the success of data science in applications where Internet-scale data is available, e.g., natural language processing, optical character recognition, object tracking, and most recently, autonomous driving, there is a growing anticipation of similar accomplishments in scientific disciplines [10], [11], [12]. To capture this excitement, some have even referred to the rise of data science in scientific disciplines as “the end of theory” [13], the idea being that, the increasingly large amounts of data makes it possible to build actionable models without using scientific theories.

Unfortunately, this notion of black-box application of data science has met with limited success in scientific domains (e.g. [14], [15], [16]). A well-known example of the perils in using data science methods in a theory-agnostic manner is Google Flu Trends, where a data-driven model was learned to estimate the number of influenza-related physician visits based on the number of influenza-related Google search queries in the United States [17]. This model was built using search terms that were highly correlated with the flu propensity in the Center for Disease Control (CDC) data. Despite its initial success, this model later overestimated the flu propensity by more than a factor of two, as measured by the number of influenza-related doctor visits in subsequent years, according to CDC data [15].

There are two primary characteristics of knowledge discovery in scientific disciplines that have prevented data science models from reaching the level of success achieved in commercial domains. First, scientific problems are often under-constrained in nature as they suffer from paucity of representative training samples while involving a large number of variables that interact in complex and non-stationary ways. For example, in the discovery of mate-

- A. Karpatne is with the University of Minnesota.
E-mail: karpa009@umn.edu
- G. Atluri is with the University of Cincinnati.
- J. Faghmous is with the Icahn School of Medicine at Mount Sinai.
- M. Steinbach, A. Banerjee, S. Shekhar, and V. Kumar are with the University of Minnesota.
- A. Ganguly is with the Northeastern University.
- N. Samatova is with the North Carolina State University.

rials, the number of materials that have been thoroughly characterized to date remains small, while the possible combinations of compounds and molecules are enormously large [18]. Note that a model that is purely data-driven can only be as good as the quality of the available data set. Indeed, a black-box model may capture specific nuances in the data arising due to sampling biases in the labeled data (available for training or cross-validation), leading to poor generalization performance. This is usually not the case with Internet-scale applications such as text mining or object recognition, where large volumes of labeled or unlabeled data are easily available.

The second difference in scientific domains is in the nature of the goal of knowledge discovery. While a common end-goal of data science models is the generation of actionable models, the process of knowledge discovery in scientific domains does not end at that. While a model that provides better movie recommendations on Netflix can be judged solely on the basis of predictive performance, the parameters of success are more elaborate in scientific problems. In particular, the *scientific interpretability* of the model can be at least as important as improving predictive performance. As an example in bio-marker discovery, it is not sufficient to simply detect the genetic traits responsible for a certain disease but it is also necessary to discover the biological mechanisms or pathways through which they affect the occurrence of the disease. Scientific interpretability is further important to safeguard against the possibility of learning spurious patterns purely from the data, especially in problems that are of critical nature and are associated with high risks. In particular, a data science model that is *consistent* with domain knowledge is less likely to show misleading performance on the labeled set that does not generalize well outside the available data.

The limitations of black-box data science models in scientific disciplines motivate a novel paradigm that uses the unique capability of data science models to automatically learn patterns and models from large data, without ignoring the treasure of accumulated scientific knowledge. We refer to this paradigm that attempts to integrate scientific knowledge and data science as *theory-guided data science* (TGDS). The paradigm of TGDS has already begun to show promise in scientific problems from diverse disciplines. Some examples include the discovery of novel climate patterns and relationships [19], [20], closure of knowledge gaps in turbulence modeling efforts [21], [22], discovery of novel compounds in material science [23], [24], [25], design of density functionals in quantum chemistry [26], improved imaging technologies in bio-medical science [27], [28], discovery of genetic biomarkers [29], and the estimation of surface water dynamics at a global scale [30], [31]. These efforts have been complimented with recent review papers [8], [18], [32], [33], workshops (e.g. a 2016 conference on physics informed machine learning [34]) and industry initiatives (e.g. a recent IBM Research initiative on “physical analytics” [35]).

This paper attempts to build the foundations of theory-guided data science by presenting several ways of bringing scientific knowledge and data science models together, and illustrating them using examples of applications from diverse domains. A major goal of this article is to formally conceptualize the paradigm of “theory-guided data science”,

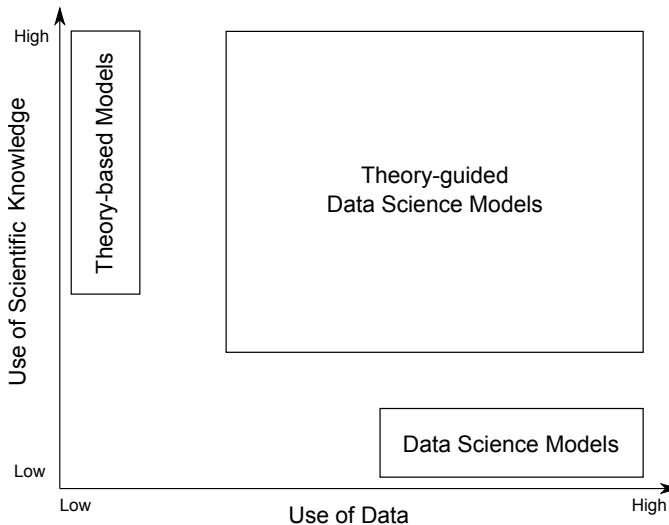


Fig. 1: A representation of knowledge discovery methods in scientific applications. The x -axis measures the use of data while the y -axis measures the use of scientific knowledge. Theory-guided data science explores the space of knowledge discovery that makes ample use of the available data while being observant of the underlying scientific knowledge.

where scientific theories are systematically integrated with data science models in the process of knowledge discovery.

The remainder of the article is structured as follows. Section 2 provides an introduction to theory-guided data science and presents an overview of research themes in TGDS. Sections 3, 4, 5, 6, and 7 describe several approaches in every research theme of TGDS, using illustrative examples from diverse disciplines. Section 8 provides concluding remarks.

2 THEORY-GUIDED DATA SCIENCE

A common problem in scientific domains is to represent relationships among physical variables, e.g., the combustion pressure and launch velocity of a rocket or the shape of an aircraft wing and its resultant air drag. The conventional approach for representing such relationships is to use models based on scientific knowledge, i.e., theory-based models. These models can range from solving closed-form equations (e.g. using Navier-Stokes equation for studying laminar flow) to running computational simulations of dynamical systems (e.g. the use of numerical models in climate science, hydrology, and turbulence modeling). An alternate approach is to use a set of training examples involving input and output variables for learning a data science model that can automatically extract relationships between the variables.

As depicted in Figure 1, theory-based and data science models represent the two extremes of knowledge discovery, which depend on only one of the two sources of information available in any scientific problem, i.e., scientific knowledge or data. They both enjoy unique strengths and have found success in different types of applications. Theory-based models (see top-left corner of Figure 1) are well-suited

for representing processes that are conceptually well understood using known scientific principles, e.g., in the design of satellite trajectories. On the other hand, traditional data science models mainly rely on the information contained in the data and thus reside in the bottom-right corner of Figure 1. They have a wide range of applicability in domains where we have ample supply of representative data samples, e.g., in Internet-scale problems such as text mining and object recognition.

Despite their individual strengths, theory-based and data science models suffer from certain deficiencies when applied in problems of great scientific relevance, where both theory and data are currently lacking. For example, a number of scientific problems involve processes that are not completely understood by our current body of knowledge, because of the inherent complexity of the processes. In such settings, theory-based models are often forced to make a number of simplifying assumptions about the physical processes, which not only leads to poor performance but also renders the model difficult to comprehend and analyze. We illustrate this scenario using the following example from hydrological modeling.

Example 1 (Hydrological Modeling).

One of the primary objectives of hydrology is to study the processes responsible for the movement, distribution, and quality of water across the planet. Some examples of such processes include the discharge of water from the atmosphere via precipitation, and the infiltration of water underneath the Earth’s surface, known as subsurface flow. Understanding subsurface flow is important as it is intricately linked with terrestrial ecosystem processes, agricultural water use, and sudden adverse events such as floods. However, our knowledge of subsurface flow using state-of-the-art hydrological models is quite limited [36]. This is mainly because subsurface flow operates in a regime that is difficult to measure directly using *in-situ* sensors such as boreholes. In addition, subsurface flow involves a number of complex sub-processes that interact in non-linear ways, which are difficult to encapsulate in current theory-based models [37]. Due to these challenges, existing hydrological models make use of a broad range of parameters in several weakly-informed physical equations. Thus, global hydrological models tend to show poor predictive performance in describing subsurface flow processes [38]. In addition, they also lose physical interpretability due to the large number of model parameters that are difficult to interpret meaningfully with respect to the domain.

□

If we were to apply “black-box” data science models (that are oblivious to domain theories) in scientific problems similar to the above example from hydrology, we would notice a completely different set of issues that lead to poor model performance. This is because scientific problems often involve complex spaces of hypotheses with non-stationary relationships among the variables that are difficult to capture solely from the data. Indeed, many scientific problems are highly under-constrained in nature, as they involve a large number of complex and heterogeneous

attributes with only a small number of representative samples. This makes it difficult to learn generalizable models solely from the data without accidentally picking spurious solutions, as exemplified by the limited success of black-box models in scientific domains.

Hence, neither a data-only nor a theory-only approach can be considered sufficient for knowledge discovery in complex scientific applications. Instead, there is a need to explore the continuum between theory-based and data science models, where both theory and data are used in a synergistic manner. The paradigm of *theory-guided data science* (TGDS) attempts to address the shortcomings of data-only and theory-only models by seamlessly blending scientific knowledge in data science models (see Figure 1). By integrating scientific knowledge in data science models, TGDS attempts to learn models that are consistent with scientific principles, termed as *physically consistent models*, that help in achieving better generalizability than models based purely on data.

To illustrate the role of “consistency with scientific knowledge” in ensuring better generalization performance, consider the example of learning a parametric model for a predictive learning problem using a limited supply of labeled samples. Ideally, we would like to learn a model that shows the best generalization performance over any unseen instance. Unfortunately, we can only observe the model performance on the available training set, which may not be truly representative of the true generalization performance (especially when the training size is small). In recognition of this fact, a number of learning frameworks have been explored to favor the selection of *simpler* models that may have lower accuracy on the training data (compared to more complex models) but are likely to have better generalization performance. This methodology, that builds on the well-known statistical principle of bias-variance trade-off [39], can be described using Figure 2.

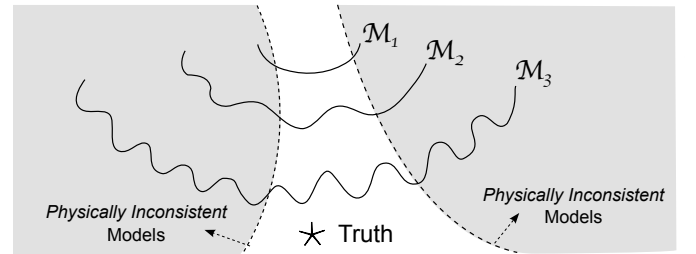


Fig. 2: Scientific knowledge can help in reducing the model variance by removing physically inconsistent solutions, without likely affecting their bias.

Figure 2 shows an abstract representation of a succession of model families with varying levels of complexity (shown as curved lines), where \mathcal{M}_1 represents the set of least complex models while \mathcal{M}_3 contains highly complex models. Every point on the curved lines represents a model that a learning algorithm can arrive at, given a particular realization of training instances. The *true* relationship between the input and output variables is depicted as a star in Figure 2. We can observe that the learned models belonging to \mathcal{M}_3 , on an average, are quite close to the true relationship. However, even a small change in the training set can bring

about large changes in the learned models of \mathcal{M}_3 . Hence, \mathcal{M}_3 shows low *bias* but high *variance*. On the other hand, models belonging to \mathcal{M}_1 are quite robust to changes in the training set and thus show low variance. However, \mathcal{M}_1 shows high bias as its models are generally farther away from the true relationship as compared to models of \mathcal{M}_3 . It is the trade-off between reducing bias and variance that is at the heart of a number of machine learning algorithms [39], [40], [41].

In scientific applications, there is another source of information that can be used to ensure the selection of generalizable models, which is the available scientific knowledge. By pruning candidate models that are inconsistent with known scientific principles (shown as shaded regions in Figure 2), we can significantly reduce the variance of models without likely affecting their bias. A learning algorithm can then be focused on the space of physically consistent models, leading to generalizable and scientifically interpretable models. This is one of the overarching visions of TGDS—to include physical *consistency* as a critical component of model performance along with training accuracy and model complexity. This can be summarized by the following revised objective of model performance:

$$\text{Performance} \propto \text{Accuracy} + \text{Simplicity} + \text{Consistency}.$$

The notion of physical consistency can be naturally incorporated in existing learning frameworks using a variety of ways. First, scientific knowledge can be used in the design of model families to restrict the space of models to physically consistent solutions, e.g., in the selection of response and loss functions or in the design of model architectures. These techniques are discussed in Section 3. Second, given a model family, we can also guide a learning algorithm to focus on physically consistent solutions. This can be achieved, for instance, by initializing the model with physically meaningful parameters, by encoding scientific knowledge as probabilistic relationships, by using domain-guided constraints, or with the help of regularization terms inspired by our physical understanding. These techniques are discussed in Section 4.

Apart from ensuring generalizability in the learning of data science models, scientific knowledge can be integrated in a variety of other ways with data science models. For example, the outputs of data science models can be refined using explicit or implicit scientific knowledge. This is discussed in Section 5. Another way of blending scientific knowledge and data science is to construct *hybrid* models, where some aspects of the problem are modeled using theory-based components while other aspects are modeled using data science components. Techniques for constructing hybrid TGDS models are discussed in Section 6. Data science methods can also help in augmenting theory-based models to make effective use of observational data. These approaches are discussed in Section 7.

3 THEORY-GUIDED DESIGN OF DATA SCIENCE MODELS

An important decision in the learning of data science models is the choice of model family used for representing the relationships between input and response variables. In

scientific applications, if the domain knowledge suggests a particular form of relationship between the inputs and outputs, care must be taken to ensure that the same form of relationship is used in the data science model. Here, we discuss two different ways of using scientific knowledge in the design of data science models. First, we can use synergistic combinations of response and loss functions (e.g. in generalized linear models or artificial neural networks) that not only simplify the optimization process and thus lead to low training errors, but are also consistent with our physical understanding and hence result in generalizable solutions. Another way to infuse domain knowledge is by choosing a model architecture (e.g. the placement of layers in artificial neural networks) that is compliant with scientific knowledge. We discuss both these approaches in the following.

3.1 Theory-guided Specification of Response

Many data science models provide the option for specifying the form of relationship used for describing the response variable. For example, a generic family of models, which can represent a broad variety of relationships between input and response variables, is the generalized linear model (GLM). There are two basic building blocks in a GLM, the link function $g(\cdot)$, and the probability distribution $P(y|\mathbf{x})$. Using these building blocks, the expected mean μ of the target variable y is determined as a function of the weighted linear combination of inputs, \mathbf{x} , as follows:

$$\begin{aligned} g(\mu) &= \mathbf{w}^T \mathbf{x} + b, \text{ or equivalently,} \\ \mu &= g^{-1}(\mathbf{w}^T \mathbf{x} + b) \end{aligned} \quad (1)$$

Some common choices of link and probability distribution functions are listed in Table 1, resulting in varying types of regression models.

To ensure the learning of GLMs that produce physically meaningful results, it is important to choose an appropriate specification of the response variable that matches with domain understanding. For example, while modeling response variables that show extreme effects (highly skewed distributions), e.g., occurrences of unusually severe floods and droughts, it would be inappropriate to assume the response variable to be Gaussian distributed (the standard assumption used in linear regression models). Instead, a regression model that uses the Gumbel distribution to model extreme values would be more accurate and physically meaningful.

One can also devise novel combinations of probability distributions and link functions to incorporate complex forms of relationships between input and response variables that are consistent with scientific knowledge. For example, let us consider the problem of predicting the average velocity V of a liquid in an open channel flow given the hydraulic

TABLE 1: Table showing some commonly used combinations of link function and probability distribution functions in generalized linear models.

Name	Link Function	Probability Distribution
Linear	μ	Gaussian
Poisson	$\log(\mu)$	Poisson
Logistic	$\log(\mu/(1 - \mu))$	Binomial

radius R . The relationship between V and R is known to follow the Manning formula [42] that can be described as:

$$V = \frac{\sqrt{S}}{n} R^{\frac{2}{3}},$$

where S is the slope and n is the Gauckler–Manning coefficient. This domain knowledge can be incorporated in the construction of a GLM for predicting V given R , with the help of the following link function:

$$\begin{aligned} V &= g^{-1}(wR) = (wR)^{\frac{2}{3}}, \\ \implies g(V) &= V^{\frac{3}{2}} = wR, \end{aligned}$$

where w is the parameter of the model that needs to be inferred. This GLM would adhere to our physical understanding of the domain while making use of the available data for training model parameters.

3.2 Theory-guided Design of Model Architecture

Scientific knowledge can also be used to influence the architecture of data science models. An example of a data science model that provides ample room for tuning the model architecture is artificial neural networks (ANN), which has recently gained widespread acceptance in several applications such as vision, speech, and language processing. There are a number of design considerations that influence the construction of an effective ANN model. Some examples include the number of hidden layers and the nature of connections among the layers, the sharing of model parameters among nodes, and the choice of activation and loss functions for effective model learning. Many of these design considerations are primarily motivated to simplify the learning procedure, minimize the training loss, and ensure robust generalization performance using statistical principles of regularization.

There is a huge opportunity in informing these design considerations with our physical understanding of a problem, to obtain generalizable as well as scientifically interpretable results. For example, in an attempt to build a model of the brain that learns view-invariant features of human faces, the use of biologically plausible rules in ANN architectures was recently explored in [43]. It was observed that along with preserving view-invariance, such theory-guided ANN models were able to capture a known aspect of human neurology (namely, the mirror-symmetric tuning to head orientation) that was being missed by traditional ANN models. This made it possible to learn scientifically interpretable models of human cognition and thus advance our understanding of the inner workings of brain. In the following, we describe two promising directions for using scientific knowledge while constructing ANN models: by using a modular design that is inspired by domain understanding, and by specifying the connections among the nodes in a physically consistent manner.

Domain knowledge can be used in the design of ANN models by decomposing the overall problem into modular sub-problems, each of which represents a different physical sub-process. Every sub-problem can then be learned using a different ANN model, whose inputs and outputs are connected with each other in accordance with the physical relationships among the sub-processes. For example,

in order to describe the overall hydrological process of surface water discharge, we can learn modular ANN models for different sub-processes such as the atmospheric process of rainfall and evaporation, the process of surface water runoff, and the process related to groundwater seepage. Every ANN model can be fed with appropriately chosen domain features at the input and output layers. This will help in using the power of deep learning frameworks while following a high-level organization in the ANN architecture that is motivated by domain knowledge.

Domain knowledge can also be used in the design of ANN models by specifying node connections that capture theory-guided dependencies among variables. A number of variants of ANN have been explored to capture spatial and temporal dependencies between the input and output variables. For example, recurrent neural networks (RNN) are able to incorporate the sequential context of time in speech and language processing [44]. RNN models have been recently explored to capture notions of long and short term memory (LSTM) with the help of skip connections among nodes to model information delay [45]. Such models can be used to incorporate time-varying domain characteristics in scientific applications. For example, while surface water runoff directly influences surface water discharge without any delay, groundwater runoff has a longer latency and contributes to the surface water discharge after some time lag. Such differences in time delay can be effectively modeled by a suitably designed LSTM model. Another variant of ANN is the convolutional neural network (CNN) [46], which has been widely applied in vision and image processing applications to capture spatial dependencies in the data. It further facilitates the sharing of model parameters so that the learned features are invariant to simple transformations such as scaling and transformation. Similar approaches can be explored to share the parameters (and thus reduce model complexity) over more generic similarity structures among the input features that are based on domain knowledge.

4 THEORY-GUIDED LEARNING OF DATA SCIENCE MODELS

Having chosen a suitable model design, the next step of model building involves navigating the search space of candidate models using a learning algorithm. In the following, we present four different ways of guiding the learning algorithm to choose physically consistent models. First, we can use physically consistent solutions as initial points in iterative learning algorithms such as gradient descent methods. Second, we can restrict the space of probabilistic models with the help of theory-guided priors and relationships. Third, scientific knowledge can be used as constraints in optimization schemes for ensuring physical consistency. Fourth, scientific knowledge can be encoded as regularization terms in the objective function of learning algorithms. We describe each of these approaches in the following.

4.1 Theory-guided Initialization

Many learning algorithms that are iterative in nature require an initial choice of model parameters as a first step to

commence the learning process. For such algorithms, an inferior initialization can lead to the learning of a poor model. Domain knowledge can help in the process of model initialization so that the learning algorithm is guided at an early stage to choose generalizable and physically consistent models.

An example of theory-guided initialization of model parameters includes a recent matrix completion approach for plant trait analysis [47], where the rows of the matrix correspond to plants from diverse environments while the columns correspond to plant traits such as leaf area, seed mass, and root length. Since observations about plant traits are sparsely available, such a plant trait matrix would be highly incomplete [48]. Filling the missing entries in a plant trait matrix can help us understand the characteristics of different plant species and their ability to adapt to varying environmental conditions. A traditional data science approach to this problem is to use matrix completion algorithms that have found great success in online recommender systems [49]. However, many of these algorithms are iterative in nature and use fixed or random values to initialize the matrix. In the presence of domain knowledge, we can improve these algorithms by using the species mean of every attribute as initial values in the matrix completion process. This relies on the basic principle that the species mean provides a robust estimate of the average behavior across all organisms. This approach has been shown to provide significant improvements in the accuracy of predicting plant traits than traditional methods [47]. Changes from the species mean can also be learned using subsequent matrix completion operations, which could be physically interpreted as the effect of varying environmental conditions on plant traits.

One of the data science models that requires special efforts in choosing an appropriate combination of initial model parameters is the artificial neural network, which is well-known to be susceptible to getting stuck at local minimas, saddle points, and flat regions in the loss curve. In the era of deep learning, much progress has been made to avoid the problem of inferior ANN initialization with the help of *pretraining* strategies. The basic idea of these strategies is to train the ANN model over a simpler problem (with ample availability of representative data) and use the trained model to initialize the learning for the original problem. These pretraining strategies have made major impact on our ability to learn complex hierarchies of features in several application domains such as speech and image processing. However, they rely on plentiful amounts of unlabeled or labeled data and hence are not directly applicable in scientific domains where the data sizes are small relative to the number of variables. One way to address this challenge is by devising novel pretraining strategies where computational simulations of theory-based models are used to initialize the ANN model. This can be especially useful when theory-based models can produce approximate simulations quickly, e.g., approximate model simulations of turbulent flow (see Example 5). Such pretrained theory-guided ANN models can then be fine-tuned using expert-quality ground truth.

4.2 Theory-guided Probabilistic Models

Probabilistic models provide a natural way to encode domain-specific relationships among variables in the form

of probabilistic graphical models, where the nodes represent variables and the edges represent interactions among the variables. Figure 3 shows an example of a graphical model for medical diagnosis, where the edges represent domain relationships between exercise, diet, and various disease symptoms.

However, this approach of manually encoding domain knowledge in graphical models requires a great deal of expert supervision, which can be cumbersome for problems involving a large number of variables with complex interactions—a common feature of scientific problems. In the presence of a large number of nodes, it is common to apply automated graph estimation techniques such as the use of graph Lasso [50]. The basic objective of such techniques is to estimate a sparse inverse covariance matrix that maximizes their likelihood given the data. To assist such techniques with scientific knowledge, a promising research direction is to explore graph estimation techniques that maximize data likelihood while limiting the search to physically consistent solutions.

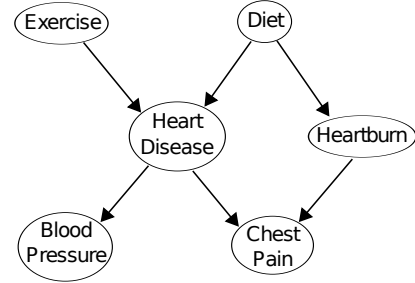


Fig. 3: An example of a Bayesian network in medical diagnosis, representing domain-driven relationships among physical variables.

Another approach to reduce the variance of model parameters (and thus avoid model overfitting) is to introduce priors in the model space. An example of the use of theory-guided priors is the problem of non-invasive electrophysiological imaging of the heart. In this problem, the electrical activity within the walls of the heart needs to be predicted based on the ECG signal measured on the torso of a subject. There are approximately 2000 locations in the walls of the heart where electrical activity needs to be predicted, based on ECG data collected from approximately 100 electrodes on the torso. Given the large space of model parameters and the paucity of labeled examples with ground-truth information, a traditional black-box model that only uses the information contained in the data is highly prone to learning spurious patterns. However, apart from the knowledge contained in the data, we also have domain knowledge about how electrical signals are transmitted to the heart via the myocardial fibre structure, which are represented using electrophysiological equations. These equations can be used to determine the spatial distribution of the electric signals in the heart at time t based on the predicted electric signals at $t - 1$. Incorporation of such theory-guided spatial distributions as a prior and using it along with externally collected ECG data in a hierarchical Bayesian model has been shown to provide promising results over traditional data science models [27], [28].

4.3 Theory-guided Constrained Optimization

Constrained optimization techniques are extensively used in data science models for restricting the space of model parameters. For example, support vector machines use constraints for ensuring separability among the classes, while maximizing the margin of the hyperplane. There is also a rich literature on constraint-based pattern mining [51], [52] and clustering [53]. The use of constraints provides a natural way to integrate domain knowledge in the learning of data science models. In scientific applications where theory-based constraints can be represented using linear equality or inequality conditions, they can be readily integrated in existing constrained optimization formulations, which are known to provide computationally efficient solutions especially when the objective function is convex.

However, many scientific problems involve constraints that are represented in complex forms, e.g., using partial differential equations (PDE) or non-linear transformations of variables, which are not easily handled by traditional constrained optimization methods. For example, the Navier-Stokes equation for momentum expresses the following constraint between the flow velocity \mathbf{v} and the fluid pressure p :

$$\rho \left(\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} \right) = -\nabla p + \nabla \cdot (\mu (\nabla \mathbf{u} + (\nabla \mathbf{u})^T)) - \frac{2}{3} \mu (\nabla \cdot \mathbf{u}) \mathbf{I},$$

where ρ is the fluid density, μ is the fluid dynamic viscosity, and ∇ represents the gradient operator with respect to the spatial coordinates.

To utilize such complex forms of constraints in data science models, it is necessary to develop constrained optimization techniques that can use common forms of partial differential equations encountered in scientific disciplines. An example of a data-driven approach that uses domain-driven PDEs can be found in a recent work in climate science [54], [55], where physically constrained time-series regression models were developed to incorporate memory effects in time as well as the nonlinear noise arising from energy-conserving interactions.

In the following, we present detailed discussions of two illustrative examples of the use of theory-guided constraints. While Example 2 explores the use of constraints for predicting electron density in computational chemistry, Example 3 explores the use of elevation-based constraints among locations for mapping surface water dynamics.

Example 2 (Computational Chemistry).

In computational chemistry, solving Schrödinger's equation is at the basis of all quantum mechanical calculations for predicting the properties of solids and molecules. Schrödinger's equation can be expressed as

$$\mathbf{H}\Psi = \mathbf{E}\Psi, \quad (2)$$

$$= (\mathbf{T} + \mathbf{U} + \mathbf{V})\Psi, \quad (3)$$

where \mathbf{H} is the electronic Hamiltonian operator, Ψ is the wavefunction that describes the quantum state of the system, and \mathbf{E} is the total energy comprising of three terms, the kinetic energy, \mathbf{T} , the electron-electron interaction energy, \mathbf{U} , and the potential energy arising due to external fields, \mathbf{V} (e.g., due to positively charged nuclei). Since the computational complexity in directly

solving the Schrödinger's equation grows rapidly with increasing number of particles, N , they are infeasible for solving large many-particle systems in practical applications.

To address this, a new class of quantum chemical modeling approaches was developed by Hohenberg and Kohn in 1964 [56], which uses the electron density $n(r)$ as a basic primitive in all calculations, instead of the wavefunction Ψ . This has resulted in the rise of density functional theory (DFT) methods, which have become a standard tool for solving many-particle systems. In DFT, every variable can be expressed as a functional of the electron density function $n(r)$ (where a functional is a function of functions). For example, the total energy \mathbf{E} can be expressed in terms of functionals of $n(r)$ as follows:

$$\mathbf{E}[n] = \mathbf{T}[n] + \mathbf{U}[n] + \mathbf{V}[n]. \quad (4)$$

The density, $n_0(r)$, that leads to the lowest total energy, $\mathbf{E}[n_0]$, is known as the ground-state density of the system, which is a critical quantity to determine.

However, obtaining $n_0(r)$ is challenging because of the interaction functional, $\mathbf{U}[n]$, whose exact form is unknown. Different approximations of the interaction term have been developed to solve for the ground-state density of a system, the most notable being the class of Kohn-Sham (KS) DFT methods. However, their performance is sensitive to the quality of approximation used in modeling the interactions. Also, KS DFT methods have a computational complexity of $O(N^3)$, which makes them challenging to apply on large systems.

To overcome the challenges in existing DFT methods, a recent work by Li et al. [26] explored the use of data science models to approximate $\mathbf{T}[n]$, and use such approximations to predict the ground-state density, $n_0(r)$. In this work, kernel ridge regression methods were used to model the kinetic energy, $\mathbf{T}[n]$, of a 4-particle system as a functional of its electron density, $n(r)$. Having learned $\hat{\mathbf{T}}[n]$, we can obtain the ground-state energy, $n_0(r)$, using the following Euler-Lagrangian equation:

$$\frac{\delta \hat{\mathbf{T}}[n_0]}{\delta n_0(r)} = \mu - v(r), \quad (5)$$

where $v(r)$ is the external potential and μ is an adjustable constant. This imposes a theory-guided constraint on the model learning, such that $\hat{\mathbf{T}}[n]$ must not only show good performance in predicting the kinetic energy, but should also be admissible to accurately estimating the ground-state density, $n_0(r)$, using Equation 5. A functional that adheres to this constraint can be called "self-consistent." It was shown in [26] that a regression model that only focuses on minimizing the training error leads to highly inconsistent solutions of the ground-state density, and is thus not useful for quantum chemical calculations. This inconsistency can be traced to the inability of regression models in capturing functional derivative forms that are used in Equation 5. In particular, the derivative of $\hat{\mathbf{T}}[n]$ can easily leave the space of densities observed in the training set, and thus arrive at ill-conditioned solutions especially when the training size is small.

To overcome this limitation, a modified Euler-Lagrange constraint was proposed in [26], which restricted the space of $n_0(r)$ to the density manifold observed in the training set. This helped in learning accurate as well as self-consistent ground-state densities using the knowledge contained in the data as well as domain theories.

□

Example 3 (Mapping Surface Water Dynamics).

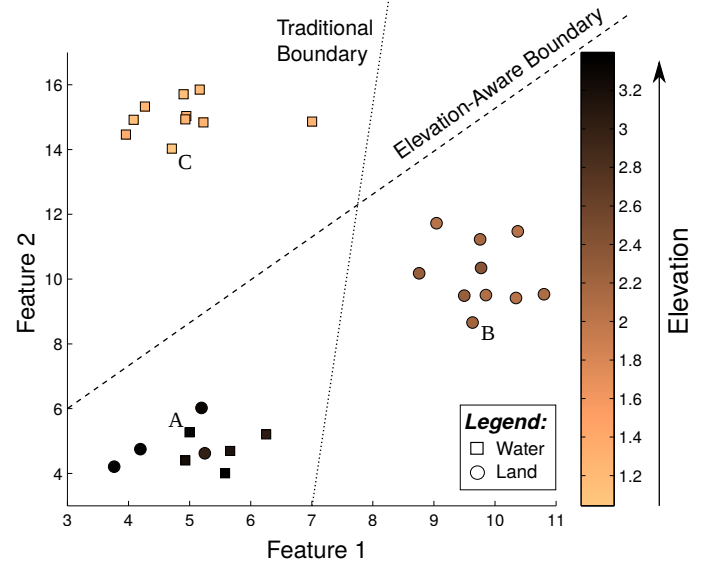
Remote sensing data from Earth observing satellites presents a promising opportunity for monitoring the extent dynamics of surface water bodies at regular intervals of time. It is possible to build predictive models that use multi-spectral data from satellite images as input features to classify pixels of the image as water or land. However, these models are challenged by the poor quality of labeled data, noise and missing values in remote sensing signals, and the inherent variability of water and land classes over space and time [57], [58].

To address these challenges, there is an opportunity for improving the quality of classification maps by using the domain knowledge that water bodies have a concave elevation structure. Hence, locations at a lower elevation are filled up first before the water level reaches locations at higher elevations. Thus, if we have access to elevation information (e.g. from bathymetric measurements obtained via sonar instruments), we can use it to constrain the classifier so that it not only minimizes the training error in the feature space but also produces labels that are consistent with the elevation structure. To illustrate this, consider an example of a two-dimensional training set shown in Figure 4a, where the squares and circles represent training instances belonging to water and land classes, respectively. Along with the features, we also have information about the elevation of every instance, shown using the intensity of colored points in Figure 4a. If we disregard the elevation information and learn a linear classifier to simply minimize the training errors, we would learn the decision boundary shown using a dotted line in Figure 4a. This classifier would make some mistakes in the lower-left corner of the feature space, where the class confusion is difficult to resolve using a linear separator. However, if we use the elevation information, we can see that the entire group of instances in the lower lower-left corner has a higher elevation than the instances shown on the right (labeled as land), and are thus less likely to be filled with water. For example, notice that location *A* is at a higher elevation than both *B* and *C* (see Figure 4b). Hence, if *B* is labeled as land, it would be inconsistent to classify *A* as water and instead it should be classified as land. The use of such constraints can help in learning a generalizable classification model even with poorly labeled training data.

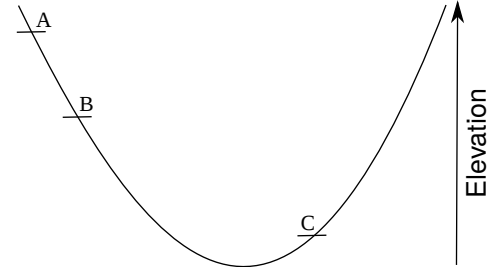
□

4.4 Theory-guided Regularization

One way to constrain the search space of model parameters is to use regularization terms in the objective function, which penalize the learning of overly complex models. A number of regularization techniques have been explored in the data science community to enforce different measures of



(a) Distribution of water and land training samples from a specific water body in feature space. Shading reflects elevation information at the locations of training samples.



(b) Lake cross-section.

Fig. 4: An illustrative example of the use of elevation-based ordering (domain theory) for learning physically consistent classification boundaries of water and land. Along with the distribution of training instances in the feature space, we also have information about their elevation, as shown in Figure 4a). This information can be used to learn an elevation-aware classification boundary that produces physically viable labels, e.g. if *B* is labeled as land, then *A* must necessarily be labeled as land as it is at a higher elevation, as shown in Figure 4b).

model complexity. For example, minimizing the L_p norm of model parameters has been extensively used for obtaining various effects of regularization in parametric model learning. While the L_2 norm has been used to avoid overly large parameter values in ridge regression and support vector machines, minimizing the L_1 norm results in the Lasso formulation and the Dantzig selector, both of which encode sparsity in the model parameters.

However, these techniques are agnostic to the physical feasibility of the learned model and thus can lead to physically inconsistent solutions. For example, while predicting the elastic modulus using bond energy and melting point, Lasso may favor melting point over bond energy even though a direct causal link exists between bond energy and the modulus [18]. This can result in the elimination of meaningful attributes and the selection of secondary at-

tributes that are not directly relevant. Hence, there is a need to devise regularization techniques that can incorporate scientific knowledge to restrict the search space of model parameters. For example, instead of using the L_p norm for regularization, we can find solutions on physically consistent sub-spaces of models. The Gaussian widths of such sub-spaces can be used as a regularization term in techniques such as the generalized Dantzig selector [59], [60]. In the following, we describe two research directions for theory-guided regularization that have been explored in different applications: using variants of Lasso to incorporate domain-specific structure among parameters, and the use of multi-task learning formulations to account for the heterogeneity in data sub-populations.

The group Lasso [61] is a useful variant of Lasso that has been explored in problems involving structured attributes. It assumes the knowledge of a grouping structure among the attributes, where only a small number of groups are considered relevant. As an example in bio-marker discovery, the groups of attributes may correspond to sets of bio-markers that are related via a common biological pathway. Group Lasso helps in selecting physically meaningful groups of attributes in the data science models, and various extensions of group Lasso have been explored for handling different types of domain characteristics, e.g., overlapping group Lasso [62], tree-guided group Lasso [63], and sparse group Lasso [64].

In a recent work [65], applications of sparse group Lasso were explored to model the domain characteristics of climate variables. In this work, climate variables observed over a range of spatial locations were used to predict a climate phenomenon of interest. By treating the set of variables observed at every location as a group, the use of group Lasso ensured that if a location is selected, all of the climate variables observed at that location will be used as relevant features. Such features thus represent meaningful (spatially coherent) regions in space that can be studied to identify physical pathways of relationships in climate science.

Another example of Lasso-based regularization that encodes domain knowledge can be found in the problem of discovering genetic markers for diseases. In this problem, data-driven approaches such as elastic nets are traditionally used to determine the relative importance of genetic markers in the context of a disease. However, geneticists understand that the relevant variants typically are located in close proximity on the genome sequence due to a property called linkage disequilibrium, which suggests that genetic information that is closely located travels together between generations of the population. This domain knowledge can be incorporated as a regularizer to ensure that the discovered genetic variants are typically located in close proximity on the genome. In fact, Liu and colleagues [29] introduced a smoothed minimax concave penalty to Lasso that captured squared differences in regression coefficients between adjacent SNPs to ensure that the difference in genetic effects between adjacent SNPs is small.

Domain knowledge can also be used to guide the regularization of a multi-task learning model, as explored for the problem of forest cover estimation in [66]. In the presence of heterogeneity in data sub-populations, different groups of instances in the data show different relationships between

the inputs and outputs. For example, different types of vegetation (e.g. forests, farms, and shrublands) may show varying responses to a target variable in remote sensing signals. MTL provides a promising solution to handle sub-population heterogeneity in such cases, by treating the learning at every sub-population as a different task. Further, by sharing the learning at related tasks, MTL enforces a robust regularization on the learning across all tasks, even in the scarcity of training data.

However, most MTL formulations require explicit knowledge of the composition of every task and the similarity structure among the tasks, which is not always known in practical applications. For example, the exact number and distribution of vegetation types is often unavailable, and when they are known, they are available at varying granularities [58]. In recent work [66], the presence of heterogeneity due to varying vegetation types was first inferred by clustering vegetation time series, which was then used to induce similarity in the model parameters at related vegetation types. This resulted in an MTL formulation where the task structure was inferred using contextual variables, obtained using domain knowledge.

5 THEORY-GUIDED REFINEMENT OF DATA SCIENCE OUTPUTS

Domain knowledge can also be used to refine the outputs of data science models so that they are in compliance with our current understanding of physical phenomena. This style of TGDS leverages scientific knowledge at the final stage of model building where the outputs of any data science model are made consistent with domain knowledge. In the following, we describe some of the approaches for refining data science outputs using domain knowledge that is either explicitly known (e.g. in the form of closed-form equations or model simulations) or implicitly available (e.g. in the form of latent constraints).

5.1 Using Explicit Domain Knowledge

Data science outputs are often refined to reduce the effect of noise and missing values and thus improve the overall quality of the results. For example, in the analysis of spatio-temporal data, there is a vast body of literature on refining model outputs to enforce spatial coherence and temporal smoothness among predictions. Data science outputs can also be refined to improve a quality measure, e.g., in the discovery of frequent itemsets by *pruning* candidate patterns. Building on these methods, there is a promise in developing model refinement approaches that make ample use of domain knowledge, encoded in the form of scientific theories, for producing physically consistent results.

An example of theory-guided refinement of data science outputs can be found in the problem of material discovery, where the objective is to find novel materials and crystal structures that show a desirable property, e.g., their ability to filter gases or to serve as a catalyst. Traditional approaches for predicting crystal structure and properties rely on *ab initio* calculations such as density functional theory methods. However, since the space of all possible materials is extremely large, it is impractical to perform computationally

expensive *ab initio* calculations on every material to estimate their structure and properties. Recently, a number of teams in material science have explored the use of probabilistic graphical models for predicting the structure and properties of a material, given a training database of materials with known structure and properties [23], [24], [25]. This provided a computationally efficient approach to reduce the space of candidate materials that show a desirable property, using the knowledge contained in the training data. The results of the data science models were then cross-checked using expensive *ab initio* calculations to further refine the model outputs. This line of research has resulted in the discovery of a hundred new ternary oxide compounds that were previously unknown using traditional approaches [23], highlighting the effectiveness of TGDS in advancing scientific knowledge.

5.2 Using Implicit Domain Knowledge

In scientific applications, the domain structure among the output variables may not always be known in the form of explicit equations that can be easily integrated in existing model refinement frameworks. This requires jointly solving the dual problem of inferring the domain constraints and using the learned constraints to refine model outputs. We illustrate this using an example in mapping surface water dynamics, where implicit constraints among locations (based on a hidden elevation ordering) are estimated and leveraged for refining classification maps of water bodies.

Example 4 (Post-processing using elevation constraints).

As described in Example 3, it is difficult to map the dynamics of surface water bodies by solely using the knowledge contained in remote sensing data, and there is a promise in using information about the elevation structure of water bodies to assist classification models. However, such information is seldom available at the desired granularity for most water bodies around the world. Hence, there is a need to infer the latent ordering among the locations (based on their elevation) so that they can be used to produce accurate and physically consistent labels. One way to achieve this is by using the history of imperfect water/land labels produced by a data science model at every location over a long period of time. In particular, a location that has been classified as water for a longer number of time-steps has a higher likelihood of being at a deeper location than a location that has been classified as water less frequently. This implicit elevation ordering, if extracted effectively, can help in improving the classification maps by post-processing the outputs to be consistent with elevation ordering. Further, the post-processed labels can help in obtaining a better estimate of the elevation ordering, thus resulting in an iterative solution that *simultaneously* infers the elevation ordering and produces physically consistent classification maps. This approach was successfully used in [30], [31] to build global maps of surface water dynamics. Figure 5 illustrates the effectiveness of this approach using an example lake in Africa, where the post-processed classification map does not suffer from the errors of the initial classification map and visually

matches well with the remote sensing image of the water body.

□

Other examples of the use of implicit constraints includes mapping urbanization [67] and tree plantation conversions [68], [69], where hidden Markov models were used to incorporate domain knowledge about the transitions among land covers.

6 LEARNING HYBRID MODELS OF THEORY AND DATA SCIENCE

One way to combine the strengths of scientific knowledge and data science is by creating *hybrid* combinations of theory-based and data science models, where some aspects of the problem are handled by theory-based components while the remaining ones are modeled using data science components. There are several ways of fusing theory-based and data science models to create hybrid TGDS models. One way is to build a two-component model where the outputs of the theory-based component are used as inputs in the data science component. This idea is used in climate science for statistical downscaling of climate variables [70], where the climate model simulations, available at coarse spatial and temporal resolutions, are used as inputs in a statistical model to predict the climate variables at finer resolutions. Theory-based model outputs can also be used to supervise the training of data science models, by providing physically consistent estimates of the target variable for every training instance.

An alternate way of creating a hybrid TGDS model is to use data science methods to predict intermediate quantities in theory-based models that are currently being missed or inaccurately estimated. By feeding data science outputs into theory-based models, such a hybrid model can not only show better predictive performance but also amend the deficiencies in existing theory-based models. Figure 6 illustrates the basic setup of this approach where the outputs of the data science model are embedded within a theory-based model. Depending on the nature of the model and the requirements of the application, there can be multiple ways of introducing data science outputs in theory-based models.

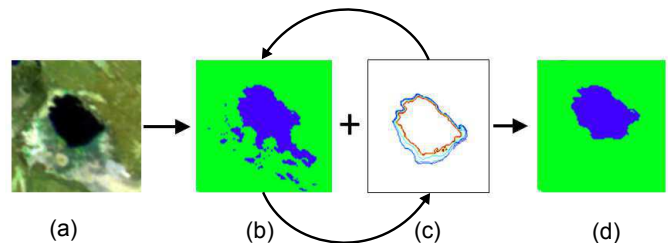


Fig. 5: Mapping the extent of Lake Abhe (on the border of Ethiopia and Djibouti in Africa) using implicit theory-guided constraints. (a) Remote sensing image of the water body (prepared using multi-spectral false color composites). (b) Initial classification maps. (c) Elevation contours inferred from the history of classification labels. (d) Final classification maps refined using elevation-based constraints.



Fig. 6: The outputs of a data science model can be chosen as intermediate quantities in a theory-based model.

In the following, we provide an illustrative example of this theme of TGDS research in the field of turbulence modeling.

Example 5 (Turbulence Modeling). One of the important problems in aerospace engineering is to model the characteristics of turbulent flow, which consists of chaotic changes in the flow velocity, and complex dissipation of momentum and energy. Turbulence modeling is used in a number of applications such as the design and reliability assessment of airfoils in aeroplanes and space vehicles. Key to the study of fluid dynamics are the Navier–Stokes equations, which describe the behavior of viscous fluids under motion. Although the Navier–Stokes equations can be readily applied in simple flow problems involving incompressible and irrotational flow, obtaining an exact representation for turbulent flow requires computationally expensive solutions such as direct numerical simulations (DNS) at fine spatial grids. The high computational costs of DNS makes it infeasible for studying practical turbulence problems in the industry, which are typically solved using inexact but computationally cheap approximations. One such approximation is the Reynolds-averaged Navier–Stokes (RANS) equations, which introduces a term called as the Reynolds stress, τ , to represent the apparent stress due to fluctuations caused by turbulence. Since the exact form of the Reynolds stress is unknown, different approximations of τ have been explored in previous studies, resulting in a variety of RANS models. Despite the continued efforts in approximating τ , current RANS models are still insufficient for modeling complex flows with separation, curvature, or swirling. To overcome their limitations, a recent work by Wang et al. [22] explored the use of machine learning methods to assist RANS models and reduce their discrepancies. In particular, the Reynolds stress was approximated as

$$\tau = \tau_{RANS} + \Delta\tau_{ML}, \quad (6)$$

where τ_{RANS} is obtained from a RANS model while $\Delta\tau_{ML}$ is the model discrepancy that is estimated using a random forest model. Although this approach can be used with any generic RANS model to estimate its discrepancy, it does not alter the form of approximation used in obtaining τ_{RANS} , since $\Delta\tau_{ML}$ is learned independently of τ_{RANS} . In another work by Singh et al. [21], a machine learning component was used to directly augment a RANS approximation in the following manner:

$$-\tau_{ij} = 2\rho\nu S_{ij}^* - \frac{2}{3}\rho K\delta_{ij}, \quad (7)$$

$$\frac{D\nu}{Dt} = \beta \times \mathbf{P} - \mathbf{D} + \mathbf{T}, \quad (8)$$

where Equation 7 is the standard Boussinesq equation relating the Reynolds stress τ_{ij} to the effective viscosity

ν , and Equation 8 is a variant of the Spalart Allmaras model that estimates ν as a function of a machine learning term, β (learned using an artificial neural network), and other physical terms, \mathbf{P} , \mathbf{D} , and \mathbf{T} , corresponding to production, destruction, and transport processes, respectively.

Both these works illustrate the potential of coupling data science outputs with theory-based models to reduce model discrepancies in complex scientific applications. The exact choice of the data science model and its contribution to the theory-based model can be explored in future investigations. Similar lines of TGDS research can be explored in other domains where current theory-based models are lacking, e.g., hydrological models for studying subsurface flow [36].

□

7 AUGMENTING THEORY-BASED MODELS USING DATA SCIENCE

There are many ways we can use data science methods to improve the effectiveness of theory-based models. Data can be assimilated in theory-based models for improved selection of model states in numerical models. Data science methods can also help in calibrating the parameters of theory-based models so that they provide a better realization of the physical system. We describe both these approaches in the following.

7.1 Data Assimilation in Theory-based Models

One of the long-standing approaches of the scientific community for integrating data in theory-based models is to use data assimilation approaches, which has been widely used in climate science and hydrology [71]. These domains typically involve dynamical systems, such as the progression of climate phenomena over time, which can be represented as a sequence of physical states in numerical models. Data assimilation is a way to infer the most likely sequence of states such that the model outputs are in agreement with the observations available at every time-step. In data assimilation, the values of the current state are constrained to depend on previous state values as well as the current data observations. For example, if we use Gaussian distribution to model the linear transition between consecutive states, this translates to a Kalman filter. However, in general, the dependencies among the states in data assimilation methods are modeled using more complex forms of distributions that are governed by physical laws and equations. Data assimilation provides a promising step in the direction of integrating data with theory-based models so that the knowledge discovery approach relies both on scientific knowledge and observational data.

7.2 Calibrating Theory-based Models using Data

Theory-based models often involve a large number of parameters in their equations that need to be calibrated in order to provide an accurate representation of the physical system. A naïve approach for model calibration is to try out every combination of parameter values, perhaps by

searching over a discrete grid defined over the parameters, and choose the combination that shows maximum likelihood with the data. However, this approach is practically infeasible when the number of parameters are large and every parameter takes many possible values. A number of computationally efficient approaches have been explored in different disciplines for parsimoniously calibrating model parameters with the help of observational data. For example, a seminal work on model calibration in the field of hydrology is the Generalized Likelihood Uncertainty Estimation (GLUE) technique [72]. This approach models the uncertainty associated with every parameter combination using Monte Carlo approaches, and uses a Bayesian formulation to incrementally update the uncertainties as new observations are made available. At any given iteration, the parameter combination that shows maximum agreement with the observations are employed in the model, the results of which are used to update the uncertainties on the next iteration.

The problem of parameter selection has recently received considerable attention in the machine learning community in the context of multi-armed bandit problems [73], [74], [75]. The basic objective in these problems is to incrementally select parameter values so that we can *explore* the space of parameter choices and *exploit* the parameter choice that provides the maximum reward, using a limited number of observations. Variants of these techniques have also been explored to operate in settings where the parameters take continuous values instead of discrete steps [76], [77]. These techniques provide a promising direction for calibrating the high-dimensional parameters of theory-based models.

8 CONCLUSION

In this paper, we formally conceptualized the paradigm of theory-guided data science (TGDS) that seeks to exploit the promise of data science without ignoring the treasure of knowledge accumulated in scientific theories and principles. We provided a taxonomy of ways in which scientific knowledge and data science can be brought together in any application with some availability of domain knowledge. We described several approaches in every research theme of TGDS, using illustrative examples from diverse disciplines. We also discussed several avenues of novel research in TGDS that need a deeper participation of the entire community for full development.

The central motivation behind TGDS is to ensure better generalization performance of models (even when the problem is complex and data samples are under-representative) by anchoring data science algorithms with scientific knowledge. By learning scientifically interpretable models, TGDS also aims at discovering novel domain insights that are currently being ignored or overly approximated by theory-based models, thus expanding our scientific knowledge.

While the TGDS research themes covered in this paper provide a comprehensive summary of recent TGDS approaches from several disciplines, they do not exhaustively cover all possible ways theory and data science can be brought together. We anticipate the discovery of novel TGDS themes in the future that explore innovative ways of

blending theory with data science, building on the taxonomy described in this paper. For example, while most of the discussion in this paper focuses on supervised learning problems, similar TGDS research themes can be explored for other traditional tasks of data mining, machine learning, and statistics. For example, scientific knowledge can be used to extract features that capture physically meaningful characteristics, e.g., scaling, translation, and rotational invariance in turbulence modeling [78], and symmetry invariance in computational chemistry [26], [79]. The use of physical principles to constrain spatio-temporal pattern mining algorithms has been explored in [80], [81] for finding ocean eddies from satellite data. The need to explore TGDS models for uncertainty quantification is discussed in [33], for understanding and projecting climate extremes. Scientific knowledge can also be used to advance other aspects of data science, e.g., the design of scientific work-flows using domain knowledge [82], [83].

We hope that this paper is a first step in starting a conversation in the community about the role of theory-guided data science in scientific discovery and how we, data scientists, can partner in such progress. We anticipate the deep integration of theory-based and data science to become a quintessential tool for scientific discovery in future research. The paradigm of TGDS, if effectively utilized, can help us realize the vision of the “fourth paradigm” [84] in its full glory, where data serves an integral role at every step of scientific knowledge discovery.

REFERENCES

- [1] G. Bell, T. Hey, and A. Szalay, “Beyond the data deluge,” *Science*, vol. 323, no. 5919, pp. 1297–1298, 2009.
- [2] Economist, “The data deluge,” *Special Supplement*, 2010.
- [3] M. James, C. Michael, B. Brad, B. Jacques, D. Richard, R. Charles, and H. Angela, “Big data: The next frontier for innovation, competition, and productivity,” *The McKinsey Global Institute*, 2011.
- [4] A. Halevy, P. Norvig, and F. Pereira, “The unreasonable effectiveness of data,” *Intelligent Systems, IEEE*, vol. 24, no. 2, pp. 8–12, 2009.
- [5] B. P. Roe, H.-J. Yang, J. Zhu, Y. Liu, I. Stancu, and G. McGregor, “Boosted decision trees as an alternative to artificial neural networks for particle identification,” *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 543, no. 2, pp. 577–584, 2005.
- [6] D. Castelvetti et al., “Artificial intelligence called in to tackle the data deluge,” *Nature*, vol. 528, no. 7580, pp. 18–19, 2015.
- [7] P. Baldi and S. Brunak, *Bioinformatics: the machine learning approach*. MIT press, 2001.
- [8] J. H. Faghmous and V. Kumar, “A Big Data Guide to Understanding Climate Change: The Case for Theory-Guided Data Science,” *Big Data*, vol. 3, 2014.
- [9] J. H. Faghmous, V. Kumar, and S. Shekhar, “Computing and climate,” *Computing in Science & Engineering*, vol. 17, no. 6, pp. 6–8, 2015.
- [10] D. Graham-Rowe, D. Goldston, C. Doctorow, M. Waldrop, C. Lynch, F. Frankel, R. Reid, S. Nelson, D. Howe, S. Rhee et al., “Big data: science in the petabyte era,” *Nature*, vol. 455, no. 7209, pp. 8–9, 2008.
- [11] T. Jonathan, A. Gerald et al., “Special issue: dealing with data,” *Science*, vol. 331, no. 6018, pp. 639–806, 2011.
- [12] T. J. Sejnowski, P. S. Churchland, and J. A. Movshon, “Putting big data to good use in neuroscience,” *Nature neuroscience*, vol. 17, no. 11, pp. 1440–1441, 2014.
- [13] C. Anderson, “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete,” *Wired Magazine*, 2008.
- [14] P. M. Caldwell, C. S. Bretherton, M. D. Zelinka, S. A. Klein, B. D. Santer, and B. M. Sanderson, “Statistical significance of climate sensitivity predictors obtained by data mining,” *Geophysical Research Letters*, vol. 41, no. 5, pp. 1803–1808, 2014.

- [15] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The Parable of Google Flu: Traps in Big Data Analysis," *Science (New York, N.Y.)*, vol. 343, no. 6176, pp. 1203–5, Mar. 2014. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/24626916>
- [16] G. Marcus and E. Davis, "Eight (no, nine!) problems with big data," *The New York Times*, vol. 6, no. 04, p. 2014, 2014.
- [17] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009.
- [18] N. Wagner and J. M. Rondinelli, "Theory-guided machine learning in materials science," *Frontiers in Materials*, vol. 3, p. 28, 2016.
- [19] J. Kawale, S. Liess, A. Kumar, M. Steinbach, P. Snyder, V. Kumar, A. R. Ganguly, N. F. Samatova, and F. Semazzi, "A graph-based approach to find teleconnections in climate data," *Statistical Analysis and Data Mining*, vol. 6, no. 3, pp. 158–179, 2013.
- [20] J. H. Faghmous, I. Frenger, Y. Yao, R. Warmka, A. Lindell, and V. Kumar, "A daily global mesoscale ocean eddy dataset from satellite altimetry," *Scientific data*, vol. 2, 2015.
- [21] A. P. Singh, S. Medida, and K. Duraisamy, "Machine learning-augmented predictive modeling of turbulent separated flows over airfoils," *arXiv preprint arXiv:1608.03990*, 2016.
- [22] J.-X. Wang, J.-L. Wu, and H. Xiao, "Physics-informed machine learning for predictive turbulence modeling: Using data to improve rans modeled reynolds stresses," *arXiv preprint arXiv:1606.07987*, 2016.
- [23] G. Hautier, C. C. Fischer, A. Jain, T. Mueller, and G. Ceder, "Finding natures missing ternary oxide compounds using machine learning and density functional theory," *Chemistry of Materials*, vol. 22, no. 12, pp. 3762–3767, 2010.
- [24] C. C. Fischer, K. J. Tibbetts, D. Morgan, and G. Ceder, "Predicting crystal structure by merging data mining with quantum mechanics," *Nature materials*, vol. 5, no. 8, pp. 641–646, 2006.
- [25] S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, "The high-throughput highway to computational materials design," *Nature materials*, vol. 12, no. 3, pp. 191–201, 2013.
- [26] L. Li, J. C. Snyder, I. M. Pelaschier, J. Huang, U.-N. Niranjan, P. Duncan, M. Rupp, K.-R. Müller, and K. Burke, "Understanding machine-learned density functionals," *International Journal of Quantum Chemistry*, 2015.
- [27] K. C. Wong, L. Wang, and P. Shi, "Active model with orthotropic hyperelastic material for cardiac image analysis," in *Functional Imaging and Modeling of the Heart*. Springer, 2009, pp. 229–238.
- [28] J. Xu, J. L. Sapp, A. R. Dehaghani, F. Gao, M. Horacek, and L. Wang, "Robust transmural electrophysiological imaging: Integrating sparse and dynamic physiological models into ecg-based inference," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Springer, 2015, pp. 519–527.
- [29] J. Liu, K. Wang, S. Ma, and J. Huang, "Accounting for linkage disequilibrium in genome-wide association studies: a penalized regression method," *Statistics and its interface*, vol. 6, no. 1, p. 99, 2013.
- [30] A. Khandelwal, V. Mithal, and V. Kumar, "Post classification label refinement using implicit ordering constraint among data instances," in *Data Mining (ICDM), 2015 IEEE International Conference on*. IEEE, 2015, pp. 799–804.
- [31] A. Khandelwal, A. Karpatne, M. Marlier, J. Kim, D. Lettenmaier, and V. Kumar, "An approach for global monitoring of surface water extent variations using modis data," in *Remote Sensing of Environment (in review)*, 2017.
- [32] J. Faghmous, A. Banerjee, S. Shekhar, M. Steinbach, V. Kumar, A. R. Ganguly, and N. Samatova, "Theory-guided data science for climate change," *Computer*, vol. 47, no. 11, pp. 74–78, 2014.
- [33] A. R. Ganguly, E. A. Kodra, A. Agrawal, A. Banerjee, S. Boriah, S. Chatterjee, S. Chatterjee, A. Choudhary, D. Das, J. Faghmous, P. Ganguli, S. Ghosh, K. Hayhoe, C. Hays, W. Hendrix, Q. Fu, J. Kawale, D. Kumar, V. Kumar, W. Liao, S. Liess, R. Mawlagedara, V. Mithal, R. Oglesby, K. Salvi, P. K. Snyder, K. Steinhaeuser, D. Wang, and D. Wuebbles, "Toward enhanced understanding and projections of climate extremes using physics-guided data mining techniques," *Nonlinear Processes in Geophysics*, vol. 21, no. 4, pp. 777–795, 2014.
- [34] *Physics Informed Machine Learning Conference*, Santa Fe, New Mexico, 2016.
- [35] "Physical analytics, ibm research," http://researcher.watson.ibm.com/researcher/view_group.php?id=6566, accessed: 2016-10-20.
- [36] M. Ghasemizade and M. Schirmer, "Subsurface flow contribution in the hydrological cycle: lessons learned and challenges ahead a review," *Environmental earth sciences*, vol. 69, no. 2, pp. 707–718, 2013.
- [37] C. Paniconi and M. Putti, "Physically based modeling in catchment hydrology at 50: Survey and outlook," *Water Resources Research*, vol. 51, no. 9, pp. 7090–7129, 2015.
- [38] M. F. Bierkens, "Global hydrology 2015: State, trends, and directions," *Water Resources Research*, vol. 51, no. 7, pp. 4923–4947, 2015.
- [39] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics Springer, Berlin, 2001, vol. 1.
- [40] P.-N. Tan, M. Steinbach, and V. Kumar, *Intorduction to Data Mining*. Addison-Wesley, 2005.
- [41] V. N. Vapnik and V. Vapnik, *Statistical learning theory*. Wiley New York, 1998, vol. 1.
- [42] R. Manning, "On the flow of water in open channels and pipes," *Transactions of the Institution of Civil Engineers of Ireland*, vol. 20, pp. 161–207, 1891.
- [43] J. Z. Leibo, Q. Liao, F. Anselmi, W. A. Freiwald, and T. Poggio, "View-tolerant face recognition and hebbian learning imply mirror-symmetric neural tuning to head orientation," *Current Biology*, 2016.
- [44] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *Interspeech*, vol. 2, 2010, p. 3.
- [45] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *INTERSPEECH*, 2014, pp. 338–342.
- [46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [47] F. Schrodt, J. Kattge, H. Shan, F. Fazayeli, J. Joswig, A. Banerjee, M. Reichstein, G. Bönsch, S. Díaz, J. Dickie *et al.*, "Bhpmf—a hierarchical bayesian approach to gap-filling and trait prediction for macroecology and functional biogeography," *Global Ecology and Biogeography*, vol. 24, no. 12, pp. 1510–1521, 2015.
- [48] J. Kattge, S. Díaz, S. Lavorel, I. Prentice, P. Leadley, G. Bönsch, E. Garnier, M. Westoby, P. B. Reich, I. Wright *et al.*, "Try—a global database of plant traits," *Global change biology*, vol. 17, no. 9, pp. 2905–2935, 2011.
- [49] P. Melville and V. Sindhwani, "Recommender systems," in *Encyclopedia of machine learning*. Springer, 2011, pp. 829–838.
- [50] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.
- [51] J.-F. Boulicaut and B. Jeudy, "Constraint-based data mining," in *Data Mining and Knowledge Discovery Handbook*. Springer, 2005, pp. 399–416.
- [52] J. Pei and J. Han, "Constrained frequent pattern mining: a pattern-growth view," *ACM SIGKDD Explorations Newsletter*, vol. 4, no. 1, pp. 31–39, 2002.
- [53] S. Basu, I. Davidson, and K. Wagstaff, *Constrained clustering: Advances in algorithms, theory, and applications*. CRC Press, 2008.
- [54] A. J. Majda and J. Harlim, "Physics constrained nonlinear regression models for time series," *Nonlinearity*, vol. 26, no. 1, p. 201, 2012.
- [55] A. J. Majda and Y. Yuan, "Fundamental limitations of ad hoc linear and quadratic multi-level regression models for physical systems," *Discrete and Continuous Dynamical Systems B*, vol. 17, no. 4, pp. 1333–1363, 2012.
- [56] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," *Physical review*, vol. 136, no. 3B, p. B864, 1964.
- [57] A. Karpatne, A. Khandelwal, X. Chen, V. Mithal, J. Faghmous, and V. Kumar, "Global monitoring of inland water dynamics: state-of-the-art, challenges, and opportunities," in *Computational Sustainability*. Springer, 2016, pp. 121–147.
- [58] A. Karpatne, Z. Jiang, R. R. Vatsavai, S. Shekhar, and V. Kumar, "Monitoring land-cover changes: A machine-learning perspective," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 8–21, 2016.
- [59] G. M. James and P. Radchenko, "A generalized dantzig selector with shrinkage tuning," *Biometrika*, vol. 96, no. 2, pp. 323–337, 2009.
- [60] S. Chatterjee, S. Chen, and A. Banerjee, "Generalized dantzig selector: Application to the k-support norm," in *Advances in Neural Information Processing Systems*, 2014, pp. 1934–1942.

- [61] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.
- [62] L. Jacob, G. Obozinski, and J.-P. Vert, "Group lasso with overlap and graph lasso," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 433–440.
- [63] S. Kim and E. P. Xing, "Tree-guided group lasso for multi-task regression with structured sparsity," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 543–550.
- [64] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group lasso and a sparse group lasso," *arXiv preprint arXiv:1001.0736*, 2010.
- [65] S. Chatterjee, K. Steinhaeuser, A. Banerjee, S. Chatterjee, and A. R. Ganguly, "Sparse group lasso: Consistency and climate applications," in *SDM*. SIAM, 2012, pp. 47–58.
- [66] A. Karpatne, A. Khandelwal, S. Boriah, and V. Kumar, "Predictive learning in the presence of heterogeneity and limited training data," in *SDM*. SIAM, 2014, pp. 253–261.
- [67] V. Mithal, A. Khandelwal, S. Boriah, K. Steinhaeuser, and V. Kumar, "Change detection from temporal sequences of class labels: application to land cover change mapping," in *SIAM International Conference on Data mining, Austin, TX, USA*. Citeseer, 2013, pp. 2–4.
- [68] X. Jia, A. Khandelwal, J. Gerber, K. Carlson, P. West, L. Samberg, and V. Kumar, "Automated plantation mapping in southeast asia using remote sensing data," Department of Computer Science, University of Minnesota, Twin Cities, Tech. Rep. 16-029, 2016.
- [69] X. Jia, A. Khandelwal, N. Guru, J. Gerber, K. Carlson, P. West, and V. Kumar, "Predict land covers with transition modeling and incremental learning," in *SIAM International Conference on Data Mining (accepted)*, 2017.
- [70] R. L. Wilby, T. Wigley, D. Conway, P. Jones, B. Hewitson, J. Main, and D. Wilks, "Statistical downscaling of general circulation model output: a comparison of methods," *Water resources research*, vol. 34, no. 11, pp. 2995–3008, 1998.
- [71] G. Evensen, *Data assimilation: the ensemble Kalman filter*. Springer Science & Business Media, 2009.
- [72] K. Beven and A. Binley, "The future of distributed models: model calibration and uncertainty prediction," *Hydrological processes*, vol. 6, no. 3, pp. 279–298, 1992.
- [73] O. Chapelle and L. Li, "An empirical evaluation of thompson sampling," in *Advances in neural information processing systems*, 2011, pp. 2249–2257.
- [74] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 661–670.
- [75] M. Jordan and T. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.
- [76] R. Agrawal, "The continuum-armed bandit problem," *SIAM journal on control and optimization*, vol. 33, no. 6, pp. 1926–1951, 1995.
- [77] R. D. Kleinberg, "Nearly tight bounds for the continuum-armed bandit problem," in *Advances in Neural Information Processing Systems*, 2004, pp. 697–704.
- [78] J. Ling and J. Templeton, "Evaluation of machine learning algorithms for prediction of regions of high reynolds averaged navier stokes uncertainty," *Physics of Fluids (1994-present)*, vol. 27, no. 8, p. 085103, 2015.
- [79] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," *Physical review letters*, vol. 108, no. 5, p. 058301, 2012.
- [80] J. H. Faghmous, L. Styles, V. Mithal, S. Boriah, S. Liess, F. Vikebo, M. d. S. Mesquita, and V. Kumar, "Eddyscan: A physically consistent ocean eddy monitoring application," in *Intelligent Data Understanding (CIDU), 2012 Conference on*, oct. 2012, pp. 96–103.
- [81] J. H. Faghmous, H. Nguyen, M. Le, and V. Kumar, "Spatio-temporal consistency as a means to identify unlabeled objects in a continuous data field," in *AAAI*, 2014, pp. 410–416.
- [82] V. G. Honavar, "The promise and potential of big data: A case for discovery informatics," *Review of Policy Research*, vol. 31, no. 4, pp. 326–330, 2014.
- [83] Y. Gil, E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau, and J. Myers, "Examining the challenges of scientific workflows," *Ieee computer*, vol. 40, no. 12, pp. 26–34, 2007.
- [84] T. Hey, S. Tansley, K. M. Tolle *et al.*, *The fourth paradigm: data-intensive scientific discovery*. Microsoft research Redmond, WA, 2009, vol. 1.